메타데이터 기반 개인용 미디어 검색/관리 시스템

김현기, 허정, 서희철, 임수종, 황이규, 장명길
한국전자통신연구원 지식마이닝연구팀
e-mail : {hkk, jeonghur, hcseo, isj, yghwang, mgjang}@etri.re.kr

A Personal Media Search/Management System based on Metadata

Hyunki Kim, Jeong Hur, Hee-Cheol Seo, Soo-Jong Lim, Yi-Gyu Hwang, Myung-Gil Jang
Knowledge Mining Research Team, ETRI

요 약

최근 개인 컴퓨터에 저장되는 다양한 미디어 정보에 대한 검색요구가 크게 대두되면서, 다양한 데스크톱 검색 시스템이 출현하고 있다. 그러나, 기존 데스크톱 검색 시스템은 파일명, 일부 제한된 메타데이터, 콘텐츠들에 대한 키워드 기반의 검색을 수행하기 때문에 사용자의 요구에 부합하는 결과를 정확하게 제시하지 못하는 문제점이 있다. 본 논문에서는 이와 같은 문제점을 해결하기 위해서 시맨틱웹 기술을 활용하여, 온톨로지에 기반한 메타데이터를 정의하고 이를 기반으로 메타데이터간의 의미적 연관성에 기반한 시맨틱 데스크톱 검색/관리 시스템에 대해 기술한다.

1. 서론

최근 웹의 확산과 더불어 디지털카메라, 멀티미디어 모바일폰, 캠코더등과 같은 디지털 콘텐츠 생성 기기의 보급이 확대됨에 따라 컴퓨터 사용자들은 매우 다양한 형태의 정보들(문서, 음악파일, 이메일, 이미지, 동영상)을 개인용 컴퓨터에 저장하고 있다. 이와 더불어 다양한 정보가 저장되어 있는 컴퓨터로부터 기존에 저장한 정보를 검색할 필요성이 대두되기 시작하였다. 이런 요구에 발맞춰 다양한 데스크톱 검색 프로그램들이 출현하였는데, 대표적으로 구글 데스크톱, 네이버 데스크톱 등이 있다. 그러나, 기존의 데스크톱 검색기술들은 파일명, 일부 제한된 메타데이터, 미디어 콘텐츠에 대한 키워드 기반 검색이기 때문에 사용자가 요구하는 정확한 정보를 검색하기에는 부족한 점이 있다. 이런 문제점을 보완하고자 데스크톱 검색에서 다양한 메타데이터를 활용한 시맨틱 검색을 적용하는 연구가 진행되고 있다.

Mac 의 새로운 운영체제 Tiger 에 탑재된 데스크톱 검색 애플리케이션인 Spotlight [2]는 특정 파일에 포함되어 있는 다양한 메타데이터 정보(파일크기, 생성일자, 생성자 등)를 이용해 검색의 효율성과 정확성을 높이고 있다.

시맨틱웹 기술을 데스크톱에 적용한 Beagle++ [9]는 시맨틱 데스크톱이라는 새로운 기술을 소개하고 있다. Beagle++는 시스템에 저장된 다양한 파일 정보, 이메일 정보와 웹 캐시 정보를 온톨로지에 기반한 메타데이터로 정의하고, 이 메타데이터와 온톨로지를 기반으로 하는 ObjectRank 알고리즘을 이용하여, 보다 정확하게 사용자의 요구에 부합하는 정보를 제공하고자 하였다.

본 논문에서는 사용자 설문조사를 기반으로 수집한 질의를 분석하여, 사용자의 요구에 적합한 정보들에 대한 메타데이터를 정의하고 이를 기반으로 온톨로지를 구성하였다. 구축된 온톨로지를 기반으로 추론기능을 이용하여 사용자의 질의에 적합한 정보를 제공할 수 있는 시맨틱 데스크톱 검색 시스템을 개발하였다.

본 논문의 구성은 2 장에서 웹과 데스크톱 검색의 차이를, 3 장에서 시맨틱 데스크톱 시스템의 구성을, 4 장과 5 장에서는 메타데이터, 온톨로지와 자동 메타데이터 태깅을, 6 장은 색인과 검색을, 7 장은 질의 분석을 소개한다. 마지막으로 8 장에서 결론과 향후 연구 방향에 대해서 언급한다.

2. 웹 검색과 데스크톱 검색

1990 대 중반부터 웹을 기반으로 한 정보의 공유가 본격화되면서, 엄청난 정보들이 웹 환경에서 하이퍼링크를 통해 관계성을 맺고 있다. 이처럼 다양한 연결관계를 가지고 분산되어 있는 웹 정보를 사용자의 요구에 맞게 쉽게 검색하기 위한 많은 연구가 진행되었다. 연구의 성과로 구글의 PageRank [17]와 같은 아주 우수한 기술들이 개발되었고, 사용자들은 자

신의 요구에 부합하는 웹 정보를 다소 쉽게 획득할 수 있게 되었다.

2000 년 초중반부터 본격적으로 연구되기 시작한 데스크톱 검색은 웹과는 달리, 컴퓨터 내에 저장된 다양한 정보를 대상으로 하기 때문에, 웹 검색의 핵심인 링크정보가 전혀 없다. 이로 인해, 사용자의 질문에 대한 검색결과를 순위화하는데 상당한 어려움이 있다. 이를 극복하기 위해서는 개인 컴퓨터 내의 다양한 콘텍스트(Context)정보와 사용자의 컴퓨터 사용 패턴과 같은 정보를 이용한 시맨틱 관계를 설정하는 것이 핵심적인 연구대상이 되었다.

최근 시맨틱웹 기술에 대한 관심이 고조되고 있다. 시맨틱웹은 다양한 온톨로지를 기반으로, 텍스트정보를 분석하여 subject, object, property 의 RDF 트리플(Triple)로 정보를 구성하여 저장하고 온톨로지 추론 기술을 이용하여 기존의 콘텐츠들에 포함된 정보로부터 새로운 정보를 생성하여 사용자의 요구에 부합하는 정보를 제공하고자 하는 목적을 가지고 있다. 그러나 시맨틱웹 기술을 웹에 적용하기에는 웹 정보가 너무나도 광범위하고, 온톨로지 구축에도 한계가 있다. 그래서, 웹보다는 정보량이 제한되고, 온톨로지 구축에 있어서도 상당히 제한되는 데스크톱 검색에 시맨틱웹 기술을 적용하고자 하는 연구가 많이 진행되고 있다 [9,11].

3. 시맨틱 데스크톱 검색 시스템

[그림 1]은 본 논문에서 소개하는 시맨틱 데스크톱 검색 시스템의 전체 구성도이다. 본 시스템은 리눅스 환경에서 개발되었으며, 부요(Booyo) [6] 리눅스와 페도라(Fedora) 리눅스에 최적화되어 있다. 시스템은 크게 미디어 수집기, 필터 및 색인기, 질의 처리기와 사용자 인터페이스로 구성된다.

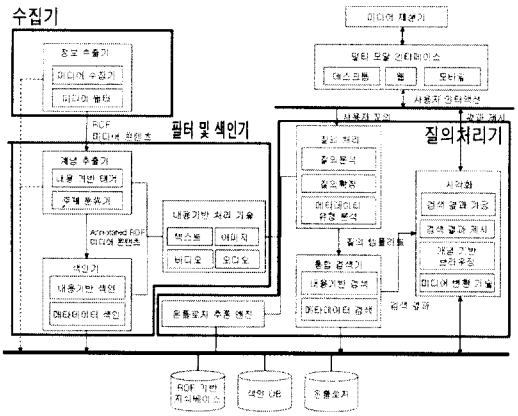

그림 1. 시맨틱 데스크톱 검색 구성도

미디어 수집기는 파일시스템의 이벤트 모니터링 매카니즘인 inotify 를 이용하여, 지정된 디렉토리 내에서 다양한 미디어 정보를 수집한다. 수집된 미디어 정보에 대해서 각 미디어 유형별로 필터를 이용하여 다양한 메타데이터를 추출한다. 특정 파일 포맷은 파일 헤더에 다양한 메타데이터 정보(EXIF [3], ID3V2 [4] 등)를 가지고 있다. 이를 추출하여, 본 시스템에서 정의한 메타데이터로 매핑한다. 또한, 내용기반 분석을 통해 메타데이터의 값을 자동으로 생성한다. 필터링과 자동메타데이터 태깅을 통해 생성된 정보는 RDF [8]로 구성되어 RDF 트리플로 저장된 후 메타데이터 필드 별 색인을 수행한다.

사용자는 사용자 GUI 를 이용하여 다양한 유형의 질의(키워드, 메타데이터, 자연어 질의)를 시스템에 입력하면, 질의처리기는 해당 질의를 분석하고, 온톨로지를 이용한 추론을 통해 사용자에게 검색결과를 제시한다. 검색결과는 다양한 시각화 기술을 이용하여 사용자에게 제시된다.

4. 메타데이터와 온톨로지

본 논문에서 메타데이터는 개인 정보와 미디어 파일 정보를 텍스트 형식으로 기술하기 위해서 사용된다. 개인 정보는 사용자와 관련된 인물 (친구, 가족 등) 정보, 장소 (여행 장소, 회의 장소 등) 정보, 이벤트 (생일, 결혼 등) 정보 등을 의미하며, 미디어 파일 정보는 개인 디지털 기기에 저장되어 있는 문서, 이미지, 오디오, 동영상 등의 파일들에 관한 정보를 의미한다.

본 논문에서 정의하는 메타데이터의 주요 목적은 미디어의 내용분석을 통한 자동 메타데이터 생성 후 미디어 검색 성능 향상에 있다. 이를 만족할 수 있는 메타데이터의 속성들을 정의하기 위해서 사용자 설문을 통한 검색 질의들을 수집하고, 수집된 검색 질의를 분석하여 메타데이터 속성을 정의하였다. 설문은 컴퓨터를 많이 사용하는 64 명을 대상으로 이루어졌으며, 설문자는 각 검색 대상 미디어 유형별로 최대 3 개의 질의를 작성하였다. 질의 분석 결과 사용자가 주로 검색하는 미디어 속성은 [표 1]과 같다. 설문자들은 이미지 검색 질의는 대체로 사진을 검색 대상으로 작성하였으며, 동영상은 영화와 자작 동영상을 함께 고려하여 질의를 작성하였다. 설문 결과에 나타난 주요 속성들을 메타데이터에 반영하였다.

표 1 질의 분석으로 수집한 주요 속성

미디어	주요 속성
문서	문서 내용, 작성 시간, 파일 확장자, 메일 보낸 사람 등
이미지	인물, 촬영 시간, 촬영 장소, 이벤트 등
오디오	가수, 장르, 제목, 발표 년도 등
동영상	이벤트, 인물, 촬영시간, 촬영장소, 제목 등

설문을 통해서 정의한 메타데이터 속성에 추가하여, 미디어 파일 정보 메타데이터에는 파일 시스템에 기술되어 있는 파일 이름, 파일 크기, 파일 생성 시간 등의 정보와 미디어 유형별로 특징적인 의미 정보가 함께 기술된다. 파일 시스템의 파일 정보는 윈도우(Microsoft Windows) 파일 시스템과 리눅스 파

일 시스템으로부터 수집하였다. 미디어 유형별 특징 정보는 각 미디어 유형별로 많이 사용되는 파일 포맷의 메타데이터를 활용하였다. 문서의 경우에는 Microsoft Office Word, Microsoft Office PowerPoint, 아래한글, 이미지의 경우에는 JPEG, EXIF, MPEG, 오디오의 경우에는 MPEG, 동영상의 경우에는 WMV, MPEG 등을 활용하였다. 그리고 추가적으로 Adobe XMP [1], vCard/vCalendar [7] 등을 고려하였다. 현재 150 여 종류의 메타데이터 속성이 정의되어 있다.

온톨로지는 메타데이터에서 사용되는 클래스와 속성 등의 정보를 계층적으로 표현한다. 온톨로지는 [그림 2]와 같은 계층구조(왼쪽)와 속성 정보(오른쪽)를 가진다. 온톨로지에는 문서, 이미지, 오디오, 동영상 등의 개인 미디어 파일에 관한 정보뿐만 아니라, 미디어 파일의 메타데이터 기술에 사용되는 인물, 장소, 범주 등의 정보가 계층적으로 표현되어 있다. 속성 정보는 미디어 파일에 대한 메타데이터 요소(element)를 구성한다. [그림 2]의 속성 정보에서 Name 이 속성, Prefix 은 속성에 대한 일종의 네임스페이스(namespace)[1], Range 는 속성 값에 대한 정보, Domain 은 속성을 가지는 클래스 정보를 의미한다. 예를 들어, 'originalArtist'라는 속성은 Prefix 는 기본값을 가지고, 속성값은 '사람' 클래스의 인스턴스(instance)이고, '오디오' 클래스의 속성이다.

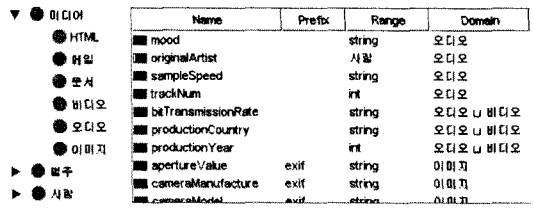

그림 2. 온톨로지 계층 구조 및 속성

온톨로지와 메타데이터는 W3C 에서 정의한 시맨틱웹 온톨로지 기술 표준 언어인 OWL (Web Ontology Language) [5]로 기술된다. OWL 과 같은 표준 언어로 기술함으로써 구축된 온톨로지와 메타데이터의 활용성이 높아질 뿐만 아니라, Protégé-OWL [16], Redland [14], Jena [10]와 같은 공개된 API 를 사용할 수 있다. 그리고 공개된 온톨로지 기반 추론 시스템인 Pellet [15], KAON2 [13] 등을 이용해서 추론이 가능하다

5. 내용기반 메타데이터 자동 태깅

본 시스템에서 정의한 메타데이터들은 크게 두 종류로 구분할 수 있다. 첫째, 기존에 특정 미디어에 존재하는 메타데이터를 활용한 것이다. 이는 필터를 통해서 메타데이터의 값을 매핑할 수 있다. 둘째, 사용자의 검색 요구를 충족하기 위해 내용기반 분석 및 사용자의 간섭에 의해 값을 채워 넣어야 하는 메타데이터이다. 대표적으로, 이미지의 컬러와 에지 디스크립션, 사람수, 실내/실외, 비디오의 키프레임 정보 등이 있다.

이미지의 컬러와 에지 디스크립션은 이미지 내용기반 검색을 위해서 자동으로 추출하는 메타데이터 값이다. 그리고, 실내/실외, 컬러/흑백, 사람수는 이미지 분석을 통해서 자동으로 값을 채워 넣는다. [표 2]는 이미지에 대한 메타데이터 자동 태깅의 정확률을 보여주고 있다.

표 2. 이미지 메타데이터 자동 태깅 정확률

	사람수	실내/실외	흑백/컬러
정확률	47%	88%	88%

비디오와 오디오의 경우, 일반적으로 관련 핵심정보(제목, 작성자 등)를 파일명으로 이용한다. 이를 이용하여, 웹으로부터 크롤링하여 구축한 영화와 음악에 대한 DB 에 파일명을 분석한 결과를 질의로 입력하여 관련 정보를 검색하고 이를 기반으로 메타데이터 값을 자동으로 태깅한다.

6. 필드기반 색인/검색

일반적인 텍스트 검색과 달리 메타데이터 별로 특성을 갖는 데이터를 색인하기 위해서 필드별로 색인하며, 온톨로지 기반 추론을 지원하기 위해 RDF 트리플을 Redland [14]에 저장한다.

메타데이터의 인스턴스는 텍스트의 본문을 포함하여 다양한 문자열로 구성이 되기 때문에 형태소 분석, 개체명 인식, 청킹을 사용한 분석한 결과를 색인하거나 이러한 분석이 필요 없을 경우 문자열 자체를 색인하였다.

검색에 이용되는 필드는 자체적으로 의미를 갖는 필드명을 사용하였지만 검색에서 지정하여 사용하지 않는 메타데이터는 공통 필드로 색인하여 필드를 지정하지 않는 검색에 대비하였다.

검색시에는 사용자가 필드를 지정할 수도 있고 지정하지 않을 경우에는 디폴트 필드를 사용하는데 본 시스템에서는 디폴트 필드로 공통 필드인 "contents" 필드를 사용하였다.

7. 질의 분석

개인용 미디어 검색에서 사용자의 질의를 분석하기 위해 자연어 형태의 사용자 질문을 수집한 결과 아래와 같은 다양한 유형이 모아졌다.

어제 작성한 doc 파일
2004 년도 마케팅 관련 보고서
어제 교수님에게서 온 메일
ppt 파일이 첨부된 메일
작년 5 월 에버랜드 사진
강릉갔을 때 찍은 사진
이승철이 부른 노래
발라드 곡을 찾아라
와이프랑 같이 찍은 동영상 검색
작년 결혼식 영상

[1] Prefix 값이 없는 속성은 기본 네임스페이스(default namespace)를 가진다.

사용자의 질문에는 찾고자 하는 미디어의 유형과 메타데이터를 이용하여 검색된 결과를 제약하는 정보 등이 포함되어 있다. 이를 분석하기 위해 두 단계로 사용자의 질의를 분석한다.

첫 번째 단계는 사용자가 찾고자 하는 미디어의 유형(문서, 이미지, 오디오, 비디오)을 인식하는 것이다. 이 단계에서는 사용자가 구체적인 미디어의 확장자가 포함되면 이것도 인식한다. 예를 들어, "워드 문서"나 "jpg 파일"과 같이 구체적인 미디어의 확장 유형을 언급한 경우, 이를 인식한다.

두 번째 단계는 사용자의 질문에 포함된 메타데이터를 인식한다. "작년에 만든 한글 문서"나 "아빠와 찍은 사진", "제주도에서 찍은 비디오" 등에서는 각각 '작년', '아빠', '제주도' 등이 미디어에 태깅된 메타데이터와 밀접한 관련이 있다. 이는 한국어 개체명(Named Entity) 인식기를 이용하여 찾아 낸다. 한 개체명 유형이 특정 미디어에서는 특정한 메타데이터로 사상될 수 있지만 반드시 그런 것은 아니다. 메일과 같은 경우, "팀장님이 보낸 메일"과 "팀장님에게 보낸 메일"은 직위 개체명 '팀장'은 같아도 어휘 패턴에 따라 수신자와 발신자로 구분될 수 있다.

인식된 개체명이 해당 미디어의 어떤 메타데이터와 관련이 있는지를 분석하기 위해 LSP (Lexico-Semantic Pattern) 기반의 100 여개 질의 패턴을 구축하고, 사용자 질문을 LSP 패턴으로 변환한 후, 구축된 패턴과 비교하는 방법을 사용하였다.

```
(&MER_PS_NAME)(%jc)*(%찍).*(%사진)
(&MER_PS_NAME).*(%jcs)(%작성).*(%문서|%파일|%
파일)
(&MER_GENRE).*(%검색)
```

정규표현식에 기반하여 어휘와 품사 및 개체명으로 구성된 LSP 의 예는 위와 같다. 질의 패턴에서 '&'는 개체명, '%'는 품사 또는 어휘를 나타낸다.

질의 분석 결과에 대해서 온톨로지 기반 추론을 통해서 질의확장을 시도한다. 질의 확장은 질의 분석 결과인 질의어에 대해서 수행되며, 추론을 통해서 얻어지는 질의어의 포함 관계(Subsumption) 정보를 질의 확장에 사용한다.

8. 결론 및 향후연구

앞서 언급한 바와 같이 본 논문에서는 리눅스 환경에 기반하여 개인 미디어를 관리하기 위한 시맨틱 데스크톱 검색 시스템에 대해서 언급하였다. 각 미디어의 유형별로 다양한 메타데이터를 정의하였고, 메타데이터 값들의 의미적 연관성을 위하여 온톨로지를 활용하였다. 각 미디어 별로 메타데이터 태깅을 통해 생성된 메타데이터 값은 RDF 트리플로 변환하여 저장하고 미디어의 콘텐츠 정보를 메타데이터 별로 색인하였다. 색인된 정보는 질의에 의한 검색결과를 통해 원하는 정보를 찾을 수도 있고, 온톨로지의 개념을 이용한 Facet 기반 브라우징을 통해서도 정보를 찾을 수도 있다. 또한, 메타데이터에 기반하여 시각화 도구를 이용한 브라우징을 통해서 정보를 검색할 수도 있다. [그림 3]은 시각화 도구인 하이퍼블릭 트리와 맵뷰에 대한 화면이다.

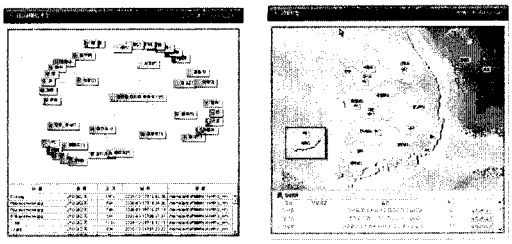

그림 3. 하이퍼블릭 트리 뷰와 맵 뷰

질의처리는 다양한 자연어 질의뿐만 아니라, 키워드 기반의 질의를 처리할 수 있도록 하였다. 질의처리에서는 온톨로지를 이용한 추론을 수행하여 의미기반 정보검색을 수행하였다.

향후 연구방향들로써는 다양한 개인들의 정보를 포괄할 수 있는 온톨로지를 구성하기란 쉽지 않기 때문에 모든 개인들이 공통적으로 인식하는 수준의 코어 온톨로지를 구성하고, 개인적인 정보와 관련된 온톨로지는 개인이 편집할 수 있도록 하는 방법이 연구되어야 할 것이다. 또한 사용자에게 의미있는 메타데이터를 다양한 포맷의 미디어에 대해 자동으로 정확하게 생성하고, 사용자 행위에 기반한 개인화된 지능형 검색 모델에 대한 연구가 필요하다.

참고문헌

[1] http://www.adobe.com/products/xmp/
[2] http://www.apple.com/macosx/features/spotlight/
[3] http://www.exif.org/
[4] http://www.id3.org/
[5] http://www.imc.org/pdi/
[6] http://www.oss.or.kr/booyo/
[7] http://www.w3.org/2004/OWL/
[8] http://www.w3.org/RDF/
[9] http://beagle.kbs.uni-hannover.de/
[10] http://jena.sourceforge.net/
[11] http://nepomuk.semanticdesktop.org/xwiki/bin/Main1/
[12] http://sourceforge.net/projects/clucene/
[13] B. Motik and U. Sattler, "Practical DL Reasoning over Large ABoxes with KAON2", submitted for publication, 2006.
[14] Davidd Beckett, "The Desing and Implementation of the Redland RDF Application Framework", The Tenth International World Wide Web Conference, Hong Kong, 2001.
[15] Evren Sirin and Bijan Parsia and Bernardo Cuenca Grau and Aditya Kalyanpur and Yarden Katz, "Pellet: a practical owl-dl reasoner", submitted for publication to Journal of Web Semantics.
[16] Holger Knublauch, Ray W. Fergerson, Natalya F. Noy, and Mark A. Musen, "The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications", Third International Semantic Web Conference – ISWC 2004, Japan (2004).
[17] Sergey Brin and Lawrence Page, "The anatomy of a large-scale hypertextual Web search engine", Proceedings of the seventh international conference on World Wide Web 7, 107-117 (1998)